



Quantile and probability-level criteria for nonlinear experimental design

Andrej Pazman, Luc Pronzato

► To cite this version:

Andrej Pazman, Luc Pronzato. Quantile and probability-level criteria for nonlinear experimental design. 8th Int. Workshop in Model-Oriented Design and Analysis, Jun 2007, Almagro, Spain. pp.157-164. hal-00170634

HAL Id: hal-00170634

<https://hal.science/hal-00170634>

Submitted on 10 Sep 2007

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Quantile and probability-level criteria for nonlinear experimental design*

Andrej Pázman¹ and Luc Pronzato²

¹ Department of Applied Mathematics and Statistics, Faculty of Mathematics,
Physics and Informatics, Comenius University, Bratislava, Slovakia
`pazman@center.fmph.uniba.sk`

² Laboratoire I3S, CNRS/Université de Nice–Sophia Antipolis, France
`pronzato@i3s.unice.fr`

Summary. We consider optimal experimental design for parameter estimation in nonlinear situations where the optimal experiment depends on the value of the parameters to be estimated. Setting a prior distribution for these parameters, we construct criteria based on quantiles and probability levels of classical design criteria and show how their derivatives can easily be approximated, so that classical algorithms for local optimal design can be used for their optimisation.

Key words: robust design, minimax optimal design, average optimal design, quantiles, nonlinear models

1 Introduction

Classical criteria for optimum experimental design for parameter estimation are functions of a normalized information matrix, which generally takes the form

$$\mathbf{M}(\xi, \theta) = \int_{\mathcal{X}} \mathcal{M}(x, \theta) \xi(dx) \quad (1)$$

with $\theta \in \mathbb{R}^p$ the parameters of interest and ξ the design, that is, a probability measure on some given region \mathcal{X} of \mathbb{R}^q . Typically, in nonlinear situations the $p \times p$ matrix $\mathcal{M}(x, \theta)$ depends on θ to be estimated. For instance, a design ξ_D is D-optimal for LS estimation in the nonlinear regression model with scalar observations $Y_k = \eta(x_k, \theta) + \varepsilon_k$ and i.i.d. errors ε_k with zero mean and finite

* The research of the first author has been supported by the VEGA grant No. 1/3016/06. The work of the second author was partially supported by the IST Programme of the European Community, under the PASCAL Network of Excellence, IST-2002-506778. This publication only reflects the authors' view.

variance ($k = 1, 2, \dots$) when it maximizes $\log \det \mathbf{M}(\xi, \theta)$ with $\mathcal{M}(x, \theta)$ the rank-one matrix

$$\mathcal{M}(x, \theta) = \frac{\partial \eta(x, \theta)}{\partial \theta} \frac{\partial \eta(x, \theta)}{\partial \theta^\top}. \quad (2)$$

Classically, a prior guess $\hat{\theta}^0$ for θ is used to design the experiment, with the hope that the local optimal design for $\hat{\theta}^0$ will be close to the optimal one for the unknown θ . When the alternation of estimation and design phases is possible, sequential design permits to progressively adapt the experiment to an estimated value of θ that (hopefully) converges to its unknown true value, see e.g. Wu (1985); Chaudhuri and Mykland (1993) for maximum-likelihood and Spokoinyi (1992); ? for Bayesian estimation. In many circumstances, however, the repetition of experimentation phases is impossible, and a single design ξ^* must be determined, based on the prior information available. Two types of approaches have been suggested to achieve some robustness with respect to a misspecification of θ . Let $\Phi(\xi, \theta)$ denote the criterion to be maximized with respect to ξ , for instance, $\Phi(\xi, \theta) = \Psi[\mathbf{M}(\xi, \theta)]$ with $\mathbf{M}(\xi, \theta)$ the information matrix (1) and $\Psi(\cdot)$ a concave function on the space of non-negative definite $p \times p$ matrices, with $p = \dim(\theta)$. *Average optimal design* puts a prior probability measure π on θ and maximizes

$$\Phi_A(\xi) = \mathbb{E}_\pi\{\Phi(\xi, \theta)\} = \int_\Theta \Phi(\xi, \theta) \pi(d\theta)$$

with $\Theta \subset \mathbb{R}^p$ the support of π , see, e.g., Fedorov (1980); Chaloner and Larntz (1989); Chaloner and Verdinelli (1995). In *maximin-optimal design* $\Phi(\xi, \theta)$ is replaced by its worst possible value for θ in Θ and the criterion to be maximized is

$$\Phi_M(\xi) = \min_{\theta \in \Theta} \Phi(\xi, \theta),$$

see, e.g., Melas (1978); Fedorov (1980); Müller and Pázman (1998). Compared to local design, average and maximin optimal design do not create any special difficulties (other than heavier computations) for discrete designs of the form $\xi = (1/n) \sum_{i=1}^n \delta_{x_i}$, with δ_x the delta measure that puts mass 1 at x and n fixed (usually, algorithms for discrete design do not exploit any special property of the design criterion, but only yield local optima). For computational reasons, the situation is simpler when π is a discrete measure and Θ is a finite set (however, a relaxation algorithm is suggested in (Pronzato and Walter, 1988) for maximin-optimal design when Θ is a compact set, and stochastic approximation can be used for average-optimal design in general situations, see, e.g., Pronzato and Walter (1985)). When optimizing a design measure (approximate design theory), the concavity of Φ is preserved, which yields Equivalence Theorems, and globally convergent algorithms can be constructed, see, e.g., Fedorov and Hackl (1997). Although attractive, average and maximin optimal design nevertheless raise several important difficulties among which are the following.

(i) A design ξ_A^* optimal for Φ_A can perform poorly for “many” values of θ , in the sense that $\pi\{\Phi(\xi_A^*, \theta) < u\}$ may be larger than α for some unacceptably low value for u and high level α .

(ii) For $g(\cdot)$ an increasing real function, the maximization of $g[\Phi(\xi, \theta)]$ is equivalent to that of $\Phi(\xi, \theta)$, but maximizing $\mathbb{E}_\pi\{g[\Phi(\xi, \theta)]\}$ is not equivalent to maximizing $\mathbb{E}_\pi\{\Phi(\xi, \theta)\}$ in general, so that a single design criterion for local optimality yields infinitely many criteria for average optimality.

(iii) Quite often an optimal design ξ_M^* for Φ_M is such that $\min_{\theta \in \Theta} \Phi(\xi_M^*, \theta)$ is reached for θ on the boundary of Θ , which makes ξ_M^* very sensitive to the choice of Θ . Also, if Θ is taken too large, it may contain values of θ such that $\mathbf{M}(\xi, \theta)$ is singular for all ξ and an optimal design may not exist.

(iv) The maximin criterion Φ_M is not differentiable everywhere, which induces some difficulties for its optimisation; in particular, the steepest-ascent direction does not necessarily correspond to a one-point delta measure.

This paper suggests new stochastic design criteria based on the distribution of $\Phi(\xi, \theta)$ when θ is distributed with some prior probability measure π on $\Theta \subset \mathbb{R}^p$. In particular, we shall consider the probability levels

$$P_u(\xi) = \pi\{\Phi(\xi, \theta) \geq u\} \quad (3)$$

and the quantiles

$$Q_\alpha(\xi) = \max\{u : P_u(\xi) \geq 1 - \alpha\}, \quad \alpha \in [0, 1], \quad (4)$$

with u and α considered as free parameters, to be chosen by the user. When the range of possible values for Φ is known (which is the case for instance when Φ is an efficiency criterion with values in $[0, 1]$), one can specify a target level u and then maximize the probability $P_u(\xi)$ that the target is reached (or equivalently minimize the risk $1 - P_u(\xi)$ that it is not). In other situations, one can specify a probability level α that defines an acceptable risk, and maximize the value of u such that the probability that $\Phi(\xi, \theta)$ is smaller than u is less than α , which corresponds to maximizing $Q_\alpha(\xi)$. We shall assume that $\Phi[(1 - \gamma)\mu + \gamma\nu, \theta]$ is continuously differentiable in $\gamma \in [0, 1]$ for any θ and any probability measures μ, ν on \mathcal{X} such that $\mathbf{M}(\mu, \theta)$ is non degenerate. We also assume that $\Phi(\xi, \theta)$ is continuous in θ and that the measure π has a positive density on every open subset of Θ . This implies that $Q_\alpha(\xi)$ is defined as the solution in u of the equation $1 - P_u(\xi) = \alpha$, see Figure 1.

One may notice that the difficulties (i-iii) mentioned above for average and maximin optimal design are explicitly taken into account by the proposed approach: the probability indicated in (i) is precisely $1 - P_u(\xi)$ which is minimized; (ii) substituting $g[\Phi(\xi, \theta)]$ for $\Phi(\xi, \theta)$ with $g(\cdot)$ increasing leaves (3) and (4) unchanged; (iii) the role of the boundary of Θ is negligible when a small probability is attached to it (and for instance probability measures with infinite support are allowed); (iv) kernel smoothing makes P_u and Q_α differentiable, see Sect. 2. When $\Phi(\xi, \theta)$ is concave in ξ for any θ , Φ_A and Φ_M are concave. Unfortunately, P_u and Q_α are generally not, which is probably the main drawback of the approach. However, Q_α obviously satisfies the

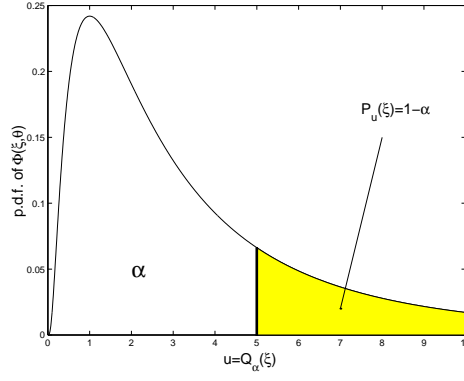


Fig. 1. Probability levels and quantiles for a design criterion $\Phi(\xi, \theta)$.

following: let Θ denote the support of π , and suppose it is compact; then, $Q_\alpha(\xi) \rightarrow \Phi_M(\xi)$ when $\alpha \rightarrow 0$, and a design optimal for Q_α will tend to be optimal for Φ_M and *vice versa*. In the next section we show how the directional derivatives of $P_u(\xi)$ and $Q_\alpha(\xi)$ can be computed, to be used in steepest-ascent optimization algorithms that converge to a local optimum (at least). An illustrative example is presented in Sect. 3 and Sect. 4 gives some conclusions and perspectives.

2 Evaluations of criteria and their derivatives

Computation of derivatives. Let $\xi = (1 - \gamma)\mu + \gamma\nu$ and consider the derivatives $\partial P_u(\xi)/\partial\gamma$ and $\partial Q_\alpha(\xi)/\partial\gamma$ at $\gamma = 0$. Since $Q_\alpha(\xi)$ satisfies the implicit equation $P_{Q_\alpha(\xi)}(\xi) = 1 - \alpha$, we can write

$$\{\partial P_u(\xi)/\partial\gamma + [\partial P_u(\xi)/\partial u][\partial Q_\alpha(\xi)/\partial\gamma]\}_{|u=Q_\alpha(\xi)} = 0,$$

which gives

$$\frac{\partial Q_\alpha(\xi)}{\partial\gamma} = - \left(\frac{\partial P_u(\xi)}{\partial\gamma} / \frac{\partial P_u(\xi)}{\partial u} \right)_{|u=Q_\alpha(\xi)}. \quad (5)$$

To compute the derivatives $\partial P_u(\xi)/\partial\gamma$ and $\partial P_u(\xi)/\partial u$ we write $P_u(\xi)$ as

$$P_u(\xi) = \int_{\Theta} \mathbf{I}_{[u, \infty)}[\Phi(\xi, \theta)] \pi(d\theta) = \int_{\Theta} \mathbf{I}_{(-\infty, \Phi(\xi, \theta)]}(u) \pi(d\theta)$$

with $\mathbf{I}_{\mathcal{A}}(\cdot)$ the indicator function of the set \mathcal{A} . When approximating the indicator step-function by a normal distribution function with small variance σ^2 , the two expressions above respectively become

$$P_u(\xi) \approx \int_{\Theta} \mathbb{F}_{u, \sigma^2}[\Phi(\xi, \theta)] \pi(d\theta) = \int_{\Theta} [1 - \mathbb{F}_{\Phi(\xi, \theta), \sigma^2}(u)] \pi(d\theta)$$

with \mathbb{F}_{a,σ^2} the distribution function of the normal $\mathcal{N}(a, \sigma^2)$. Differentiating these approximations respectively with respect to γ and u , we get

$$\frac{\partial P_u(\xi)}{\partial \gamma} \Big|_{\gamma=0} \approx \int_{\Theta} \varphi_{u,\sigma^2}[\Phi(\mu, \theta)] \frac{\partial \Phi(\xi, \theta)}{\partial \gamma} \Big|_{\gamma=0} \pi(d\theta), \quad (6)$$

$$\frac{\partial P_u(\xi)}{\partial u} \Big|_{\gamma=0} \approx - \int_{\Theta} \varphi_{\Phi(\mu, \theta), \sigma^2}(u) \pi(d\theta), \quad (7)$$

with φ_{a,σ^2} the density of \mathbb{F}_{a,σ^2} , which can be substituted in (5) to form an approximation of $\partial Q_\alpha(\xi)/\partial \gamma|_{\gamma=0}$. As shown below, this type of approximation can be related to another one, namely kernel smoothing.

Kernel smoothing. In order to estimate $P_u(\xi)$, $Q_\alpha(\xi)$ and their derivatives, one can also approximate the probability density function (p.d.f.) of $\Phi(\xi, \theta)$ by a standard kernel estimator $\phi_{n,\xi}(z) = 1/(nh_n) \sum_{i=1}^n K\left\{[z - \Phi(\xi, \hat{\theta}^i)]/h_n\right\}$. Here K is a symmetric kernel function (the p.d.f. of a probability measure on \mathbb{R} with $K(z) = K(-z)$, e.g. $\varphi_{0,1}(\cdot)$) and $\hat{\theta}^i$ ($i = 1, \dots, n$) is a sample of possible values for θ (e.g. independently randomly generated with the prior measure π). The bandwidth h_n tends to zero as $n \rightarrow \infty$. From this we obtain directly

$$P_u(\xi) \approx \hat{P}_u^n(\xi) = \int_{-\infty}^{\infty} \mathbb{I}_{[u, \infty)}(z) \phi_{n,\xi}(z) dz,$$

which is easily computed when $\int_u^\infty K(z)dz$ has a simple form. The value of $Q_\alpha(\xi)$ can then be estimated by $\hat{Q}_\alpha^n(\xi) = \{u : \hat{P}_u^n(\xi) = 1 - \alpha\}$, which is easily computed numerically. Consider now the computation of derivatives, with again $\xi = (1 - \gamma)\mu + \gamma\nu$. Direct calculations give

$$\frac{\partial \hat{P}_u^n(\xi)}{\partial \gamma} \Big|_{\gamma=0} = \frac{1}{nh_n} \sum_{i=1}^n \frac{\partial \Phi(\xi, \hat{\theta}^i)}{\partial \gamma} \Big|_{\gamma=0} K\left(\frac{u - \Phi(\mu, \hat{\theta}^i)}{h_n}\right), \quad (8)$$

$$\frac{\partial \hat{P}_u^n(\xi)}{\partial u} \Big|_{\gamma=0} = -\frac{1}{nh_n} \sum_{i=1}^n K\left(\frac{u - \Phi(\mu, \hat{\theta}^i)}{h_n}\right). \quad (9)$$

Notice that taking $\sigma^2 = h_n$ and π the discrete measure with mass $1/n$ at each $\hat{\theta}^i$ in (6, 7) respectively gives (8) and (9) with $K = \varphi_{0,1}$, the density of the standard normal. Obviously, the accuracy of these kernel approximations improves as n increases (with the only limitation due to the computational cost that increases with n).

3 Example

To illustrate the feasibility of the approach we consider D-optimal design for the nonlinear regression model $\eta(x, \theta) = \beta e^{-\lambda x}$, with $\theta = (\beta, \lambda)^\top$ the

vector of parameters to be estimated. The information matrix $\mathbf{M}(\xi, \theta)$ for a design measure ξ then takes the form (1, 2). We suppose that $\beta > 0$ and take $\mathcal{X} = [0, \infty)$. The local D-optimal experiment $\xi_D(\theta)$ that maximizes $\det \mathbf{M}(\xi, \theta)$ puts mass $1/2$ at $x = 0$ and $x = 1/\lambda$, the associated value of $\det \mathbf{M}(\xi, \theta)$ is $\det \mathbf{M}[\xi_D(\theta), \theta] = \beta^2/(4e^2\lambda^2)$. We consider the D-efficiency criterion defined by $\Phi(\xi, \theta) = \{\det \mathbf{M}(\xi, \theta)/\det \mathbf{M}[\xi_D(\theta), \theta]\}^{1/2}$, with $\Phi(\xi, \theta) \in [0, 1]$. Due to the linear dependency of $\eta(x, \theta)$ in β , $\xi_D(\theta)$ and $\Phi(\xi, \theta)$ only depend on λ and we shall simply write $\xi_D(\lambda)$, $\Phi(\xi, \lambda)$. Supposing that $\lambda = 2$ when designing the experiment, the efficiency $\Phi[\xi_D(2), \lambda]$ is plotted in solid line in Figure 2.

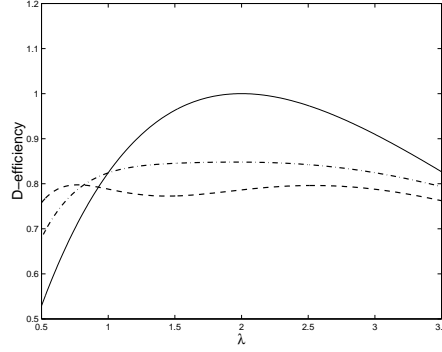


Fig. 2. D-efficiencies as function of λ for different designs; solid line: local D-optimal $\xi_D(2)$; dashed line: optimal for $P_{0.75}$; dash-dotted line optimal for $Q_{0.10}$.

Suppose now that we only know that $\lambda \in [1/2, 7/2]$ and put a uniform prior for λ on that interval; $\xi_D(2)$ is then optimal for the midpoint, but its efficiency is less than 53% for the endpoint $\lambda = 1/2$. We approximate $P_u(\xi)$ and $Q_\alpha(\xi)$ by kernel smoothing with $K = \varphi_{0,1}$ for $n = 100$ values $\hat{\lambda}^i$ equally spaced in $[0.5, 3.5]$. No special care is taken for the choice of h_n , and we simply use the rule $h_n = \hat{\sigma}_n(\Phi)n^{-1/5}$ with $\hat{\sigma}_n(\Phi)$ the empirical standard deviation of the values $\Phi(\xi, \lambda^i)$, $i = 1, \dots, n$. Figure 3 shows the estimated values \hat{P}_u^n (left) and \hat{Q}_α^n (right), in dashed lines, as functions of u and α respectively, for $\xi = \xi_D(2)$. One can check the reasonably good agreement with the exact values of P_u and Q_α , plotted in solid lines (increasing n to 1000 makes the curves almost indistinguishable).

The optimisation of $\hat{P}_{0.75}^n$ and $\hat{Q}_{0.10}^n$ with a vertex-direction (steepest-ascent) algorithm on the finite design space $\{0, 0.1, 0.2, \dots, 5\}$ respectively gives the four-point designs

$$\xi^*(P_{0.75}) \approx \begin{Bmatrix} 0 & 0.3 & 0.4 & 1.7 \\ 0.4523 & 0.0977 & 0.2532 & 0.1968 \end{Bmatrix},$$

$$\xi^*(Q_{0.10}) \approx \begin{Bmatrix} 0 & 0.3 & 0.4 & 1.3 \\ 0.4688 & 0.1008 & 0.2634 & 0.1670 \end{Bmatrix},$$

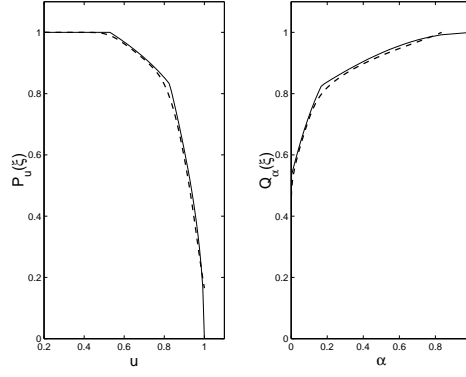


Fig. 3. Left: $\hat{P}_u^n(\xi)$ (dashed line) and $P_u(\xi)$ (solid line) as functions of u ; right: $\hat{Q}_\alpha^n(\xi)$ (dashed line) and $Q_\alpha(\xi)$ (solid line) as functions of α ; $\xi = \xi_D(2)$, $n = 100$.

where the first row indicates the support points and the second one their respective weights. They satisfy $\hat{P}_{0.75}^n[\xi^*(P_{0.75})] \approx 0.9999$ and $\hat{Q}_\alpha^n[\xi^*(Q_{0.10})] \approx 0.783$. The efficiencies of these designs are plotted in Figure 2. The exact value $P_u[\xi^*(P_{0.75})]$ equals one, indicating that the efficiency is larger than 75% for all possible values of λ . The optimisation of $\hat{Q}_{0.01}^n$ gives a design very close to $\xi^*(P_{0.75})$ which, together with the shape of the curve in dashed line on Figure 2, suggests that $\xi^*(P_{0.75})$ is almost maximin optimal. Accepting a small loss of efficiency for about 10% of the values of λ produces a significant increase of efficiency on most of the interval, see the curve in dash-dotted line.

4 Conclusions and further developments

The paper shows the feasibility of optimal design based on quantiles and probability level criteria in the situations where the local optimal experiment depends on the unknown parameters to be estimated. In particular, kernel smoothing permits to optimise design measures with classical algorithms borrowed from local optimal design. Adapting the sample size n , the kernel K and the bandwidth h_n to a particular problem, and maybe a particular algorithm, may deserve further studies. In particular, one might think of letting n grow with the number of iterations of the algorithm, as in stochastic approximation methods, see, e.g., Chapter 4 of (Kibzun and Kan, 1996).

Notice, finally that the ideas presented in the paper are very general and could also be applied to discrete design based on more accurate descriptions of parameter uncertainty than functions of information matrices, such as the volumes of confidence regions (Hamilton and Watts, 1985), the mean-squared error (Pázman and Pronzato, 1992; Gauchi and Pázman, 2006) or the entropy of the distribution of the LS estimator (Pronzato and Pázman, 1994).

References

- K. Chaloner and K. Larntz. Optimal Bayesian design applied to logistic regression experiments. *Journal of Statistical Planning and Inference*, 21: 191–208, 1989.
- K. Chaloner and I. Verdinelli. Bayesian experimental design: a review. *Statistical Science*, 10(3):273–304, 1995.
- P. Chaudhuri and P.A. Mykland. Nonlinear experiments: optimal design and inference based likelihood. *Journal of the American Statistical Association*, 88(422):538–546, 1993.
- V.V. Fedorov. Convex design theory. *Math. Operationsforsch. Statist., Ser. Statistics*, 11(3):403–413, 1980.
- V.V. Fedorov and P. Hackl. *Model-Oriented Design of Experiments*. Springer, Berlin, 1997.
- J.-P. Gauchi and A. Pázman. Designs in nonlinear regression by stochastic minimization of functionals of the mean square error matrix. *Journal of Statistical Planning and Inference*, 136:1135–1152, 2006.
- D.C. Hamilton and D.G. Watts. A quadratic design criterion for precise estimation in nonlinear regression models. *Technometrics*, 27:241–250, 1985.
- A.I. Kibzun and Y.S. Kan. *Stochastic Programming Problems*. Wiley, New York, 1996.
- V.B. Melas. Optimal designs for exponential regressions. *Math. Operationsforsch. und Statist., Ser. Statistics*, 9:753–768, 1978.
- Ch.H. Müller and A. Pázman. Applications of necessary and sufficient conditions for maximin efficient designs. *Metrika*, 48:1–19, 1998.
- A. Pázman and L. Pronzato. Nonlinear experimental design based on the distribution of estimators. *Journal of Statistical Planning and Inference*, 33:385–402, 1992.
- L. Pronzato and A. Pázman. Second-order approximation of the entropy in nonlinear least-squares estimation. *Kybernetika*, 30(2):187–198, 1994. *Erratum* 32(1):104, 1996.
- L. Pronzato and E. Walter. Robust experiment design via stochastic approximation. *Mathematical Biosciences*, 75:103–120, 1985.
- L. Pronzato and E. Walter. Robust experiment design via maximin optimization. *Mathematical Biosciences*, 89:161–176, 1988.
- V.G. Spokoinyi. On asymptotically optimal sequential experimental design. *Advances in Soviet Mathematics*, 12:135–150, 1992.
- C.F.J. Wu. Asymptotic inference from sequential design in a nonlinear situation. *Biometrika*, 72(3):553–558, 1985.